Genome Medicine

Minireview

Exploring the unknown: assumptions about allelic architecture and strategies for susceptibility variant discovery

Mark I McCarthy

Addresses: Oxford Centre for Diabetes, Endocrinology & Metabolism, University of Oxford, Churchill Hospital, Oxford OX3 7LJ, UK, and the Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX7 7BN, UK.
Email: mark.mccarthy@drl.ox.ac.uk

Abstract
Identification of common-variant associations for many common disorders has been highly effective, but the loci detected so far typically explain only a small proportion of the genetic predisposition to disease. Extending explained genetic variance is one of the major near-term goals of human genetic research. Next-generation sequencing technologies offer great promise, but optimal strategies for their deployment remain uncertain, not least because we lack a clear view of the characteristics of the variants being sought. Here, I discuss what can and cannot be inferred about complex trait disease architecture from the information currently available and review the implications for future research strategies.

Genome-wide association (GWA) analysis has provided the first effective strategy to allow a systematic dissection of the genetic basis of common, complex, multifactorial traits [1,2]. Several hundred loci have been identified to stringent levels of significance [3]. Although for many of these we remain some distance from a complete enumeration of causal mechanisms, there have already been substantial advances in understanding of disease - the role of autophagy in inflammatory bowel disease [4] and cell adhesion in autism [5,6], for instance.

However, for most common traits the proportion of the overall phenotypic variance explained remains small, limiting the extent to which prediction of individual disease risk is possible. There is growing speculation about the mechanisms that might account for the substantial proportion of trait heritability that remains to be characterized [7].

This speculation has repercussions well beyond recondite theoretical discussion about the genetic architecture of complex traits. With advances in technology (particularly next-generation sequencing) and growing enthusiasm for funding large-scale gene discovery efforts, hypotheses about the nature of this so-called 'genetic dark matter' [7] have a direct bearing on research strategies. Recently, this debate has seemed increasingly polarized between those who feel a continued search for common susceptibility variants is of limited value, because all that remains to be found are variants of vanishingly small effect [8], and those who feel that, pending reductions in costs that will allow high-quality, whole-genome sequence data to be generated in adequately powered sample sizes, there is virtue in persisting with an approach of proven worth [9].

There is good reason to assume that this 'dark matter' is neither an illusion created by inflated estimates of heritability nor the consequence of marked non-additivity of effects [10,11]. If so, then the sum total of genetic variance should largely be explicable in terms of the main effects of all the risk alleles of various types (single nucleotide polymorphisms, indels, copy number variants (CNVs) and inversions), allele frequencies (rare, low-frequency and common) and effect sizes. So far, the only parts of this 'space' explored systematically are those occupied by rare, penetrant alleles (principally through linkage analysis of monogenic phenotypes) and common, mostly low-effect alleles (accessible through GWA analysis). As we seek to make sensible decisions about the direction of future discovery efforts - in terms of the characteristics of the variants we are seeking and the technologies we should use to find them - we need to understand what the exploration of the 'known' genetic landscape can tell us about the parts that remain largely uncharted.

Contrasting views of the genetic landscape

One long-standing view is that complex trait susceptibility is predominantly a matter of common variants [12]. Common variants collectively account for most individual variation in DNA sequence, and the same might be expected to be the case for phenotypic variation. If true, the results of GWA studies so far indicate that most of the as-yet-undiscovered variants must (in Europeans at least) have very small effects, because the high coverage and large sample sizes used will have left few, if any, large common-variant effects undiscovered. Evidence (for example, from large-scale meta-analyses [13]) is, for many

CNV, copy number variant; GWA, genome-wide association.

traits, consistent with the notion of a long 'polygenic tail' of small effects, but it remains unclear how much of overall heritability can be explained under this model. The idea that complex-trait susceptibility involves a very large number of variants of modest effect has led some to suggest that the value of all such discoveries is diminished, on the basis that one learns little about the biology of disease if too many genes are implicated [8]. However, for many phenotypes, the overall salience of the loci of greatest effect emerging from GWA studies (the pathways implicated and the relationships to monogenic forms of the same traits) argues forcefully against such a nihilistic interpretation [9,13,14].

The contrasting viewpoint holds that common-trait susceptibility derives mostly from the action of rare or low-frequency variants [15,16]. Although such variants account for less individual sequence variation than common variants, there may be a disproportionate effect on disease susceptibility. The more recent origin of low-frequency variants may allow alleles with more dramatic phenotypic effects to be represented in the population. Also, large-effect alleles may cause phenotypic disturbances that are not as easily buffered by compensatory changes during development as are well tolerated, small-effect, common-variant alleles. Recent evidence that large, rare CNVs are associated with behavioral and psychiatric disease phenotypes [5,17,18] supports this view. Some argue that such a rare variant architecture is precisely what one would expect for diseases causing low reproductive fitness, though this rationalization fails to explain the high yield of common-variant signals reported for other diseases, such as type 1 diabetes, that were, until recently, fatal during early life [19]. It has even been suggested that many of the common-variant associations discovered by recent GWA studies may turn out to be due to the concerted action of multiple low-frequency and rare causal variants. The *NOD2* (*CARD15*) signal for Crohn's disease indicates that this is certainly possible [20]. For many diseases, however, evidence that common-variant associations are consistent across multiple ethnic groups [21] represents a strong counter to such a model: one would expect the linkage disequilibrium patterns around recent rare and low-frequency causal variants to result in far more inter-ethnic heterogeneity than is actually observed.

The best of both worlds

Although both extreme positions have merit, the likelihood is that, for most diseases, the architecture of predisposition features causal variants that have a wide range of allele frequencies and effect sizes. For most complex traits, the absence of compelling signals from linkage studies conducted in families segregating multifactorial diseases imposes an upper bound to feasible effect sizes; even so, it is easy to show that a limited number of low-frequency susceptibility alleles of medium effect could go a long way to explaining missing heritability. For example, the effect of a low-frequency variant with a population minor allele frequency of 1% and a per-allele odds ratio of 3, when measured in terms of sibling relative risk (a commonly used measure of familial aggregation), exceeds that of the largest common-variant effect known for type 2 diabetes (around *TCF7L2*). Twenty such variants across the genome would account for most of the unexplained heritability for this condition. Such a constellation of variants could provide a respectable tool for individual disease prediction, and the variants discovered would (because of their relatively large effect size) be valuable resources for detailed molecular and physiological study. The extent to which variants with these characteristics are segregating in the population remains unknown, but this is an area in which the combination of next-generation sequencing technologies and large-scale association analysis provides a powerful stimulus to discovery. Early results of this approach (such as the identification of low-frequency variants within the *IFIH1* gene that have a marked effect on type 1 diabetes susceptibility) are encouraging [22].

Strategy and the 'lumpiness' of the genome

Ultimately, we can expect large-scale, high-depth, genome-wide sequencing to enable the systematic exploration of the entire allele-frequency, effect-size space and provide empirical resolution of many of these issues. However, there remain serious financial, logistical and analytical barriers to the implementation of this technology, and the number of such experiments that could be supported by the major funders is, for the time being, limited.

All this means that, for the next few years, the power of next-generation sequencing will need to be used carefully if a profusion of underpowered discovery efforts is to be avoided. Efforts targeted to specific genomic regions (around particular candidate genes or pathways or exons across the genome, for example) are attractive because high coverage of the selected areas in large sample sizes can be generated at reasonable cost. Whole-genome sequencing will, for now, be restricted to low-pass coverage across respectable sample sizes, or high-depth coverage in smaller, highly selected, phenotypically extreme sample sets.

The genomic distribution of disease-effect loci will have a major impact on the success of these alternative approaches (Figure 1). If the low-frequency and rare variants influencing a given trait are disproportionately located in the same loci as the common variants that have been found to date, then targeted follow-up of regions revealed by GWA studies will be a powerful approach, and extending the range and scope of GWA analysis (to other ethnic groups, for example) should be a particularly efficient strategy. If, on the other hand, the 'dark matter' variants have little positional (or biological) overlap with those already known, then genome-wide resequencing is

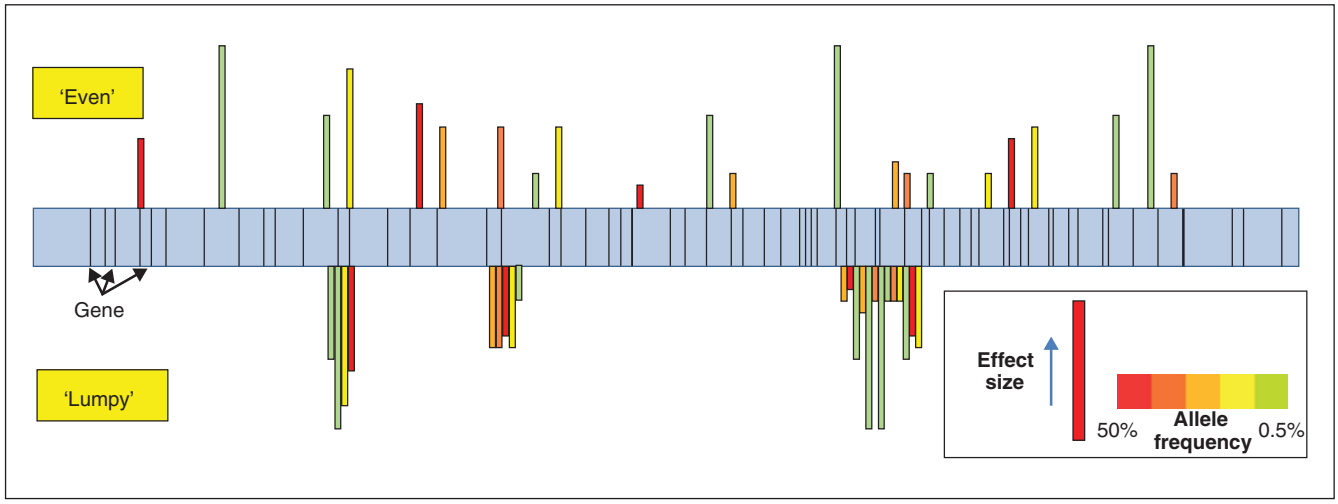

Figure 1

Causal variant signals and their genomic distribution. Two possible versions of the state of nature are presented (see text). In one ('even'), causal variants differing in terms of allele frequency (color scale) and effect size (height of bar) are distributed randomly across the genome: the location of common-variant (red/orange) associations of modest effect provides no guide to the location of lower-frequency variants (yellow/green), some of which have quite large effects. In the other ('lumpy'), causal variants congregate around certain genomic positions ('genes'): GWA studies that reveal the location of the common-variant associations will also reveal the positions of lower-frequency variants, and the proportion of disease biology explained by the loci discovered through GWA studies will be far greater than the proportion of variance explained would suggest.

likely to be the only practical way to find them. The evidence so far (overlap between monogenic and multifactorial loci; growing numbers of loci with multiple independent association signals; extensive pleiotropy, and so on [23,24]) provides some support for the former view. Effort in tracking down common susceptibility variants, as well as being valuable in its own right, should therefore guide researchers towards other types of causal variants.

Letting several well designed flowers bloom

With only limited empirical data to guide future locus-discovery efforts, extrapolation from the modest proportion of genetic variance so far explained is fraught with danger. The menu of possible research strategies is large, but each choice makes some implicit assumption about the characteristics of the variants being sought and the genomic architecture of the disease under consideration. Given uncertainties over the true state of nature, it is difficult to say which approaches will be most productive. This argues for open minds, a healthy disdain for orthodoxy, and careful exploration of the technological and methodological options. At the same time, it is important that the next wave of large-scale discovery efforts is designed so as to test assumptions about trait architecture and technological performance so that lessons of generic value to the field can be learned.

Competing interests

The author declares that he has no competing interests.

Acknowledgements

I thank the many colleagues around the world who contributed to the discussions that informed this article.

References

1. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN: **Genome-wide association studies for complex traits: consensus, uncertainty and challenges.** *Nat Rev Genet* 2008, **9:**356-369.
2. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: **Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.** *Proc Natl Acad Sci USA* 2009, **106:**9362-9367.
3. **OPG: a Catalog of Published Genome-Wide Association Studies** [http://www.genome.gov/gwastudies]
4. Cadwell K, Liu JY, Brown SL, Miyoshi H, Loh J, Lennerz JK, Kishi C, Kc W, Carrero JA, Hunt S, Stone CD, Brunt EM, Xavier RJ, Sleckman BP, Li E, Mizushima N, Stappenbeck TS, Virgin HW 4th: **A key role for autophagy and the autophagy gene *Atg16l1* in mouse and human intestinal Paneth cells.** *Nature* 2008, **456:**259-263.
5. Glessner JT, Wang K, Cai G, Korvatska O, Kim CE, Wood S, Zhang H, Estes A, Brune CW, Bradfield JP, Imielinski M, Frackelton EC, Reichert J, Crawford EL, Munson J, Sleiman PM, Chiavacci R, Annaiah K, Thomas K, Hou C, Glaberson W, Flory J, Otieno F, Garris M, Soorya L, Klei L, Piven J, Meyer KJ, Anagnostou E, Sakurai T, *et al.*: **Autism genome-wide copy number variation reveals ubiquitin and neuronal genes.** *Nature* 2009, **459:**569-573.
6. Wang K, Zhang H, Ma D, Bucan M, Glessner JT, Abrahams BS, Salyakina D, Imielinski M, Bradfield JP, Sleiman PM, Kim CE, Hou C, Frackelton E, Chiavacci R, Takahashi N, Sakurai T, Rappaport E, Lajonchere CM, Munson J, Estes A, Korvatska O, Piven J, Sonnenblick LI, Alvarez Retuerto AI, Herman EI, Dong H, Hutman T, Sigman M, Ozonoff S, Klin A, *et al.*:

Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature* 2009, **459:**528-533.

7. Maher B: **Personal genomes: the case of the missing heritability.** *Nature* 2008, **456:**18-21.
8. Goldstein DB: **Common genetic variation and human traits.** *N Engl J Med* 2009, **360:**1696-1698.
9. Hirschhorn JN: **Genome-wide association studies - illuminating biologic pathways.** *N Engl J Med* 2009, **360:**1699-1701.
10. Visscher PM, Hill WG, Wray NR: **Heritability in the genomics era - concepts and misconceptions.** *Nat Rev Genet* 2008, **9:**255-266.
11. Hill WG, Goddard ME, Visscher PM: **Data and theory point to mainly additive genetic variance for complex traits.** *PLoS Genet* 2008, **4:**e1000008.
12. Doris PA: **Hypertension genetics, single nucleotide polymorphisms, and the common disease: common variant hypothesis.** *Hypertension* 2002, **39:**323-331.
13. Weedon MN, Lango H, Lindgren CM, Wallace C, Evans DM, Mangino M, Freathy RM, Perry JR, Stevens S, Hall AS, Samani NJ, Shields B, Prokopenko I, Farrall M, Dominiczak A; Diabetes Genetics Initiative; Wellcome Trust Case Control Consortium, Johnson T, Bergmann S, Beckmann JS, Vollenweider P, Waterworth DM, Mooser V, Palmer CN, Morris AD, Ouwehand WH; Cambridge GEM Consortium, Zhao JH, Li S, Loos RJ, *et al.:* **Genome-wide association analysis identifies 20 loci that influence adult height.** *Nat Genet* 2008, **40:**575-583.
14. Loos RJF, Lindgren CM, Li S, Wheeler E, Zhao JH, Prokopenko I, Inouye M, Freathy RM, Attwood AP, Beckmann JS, Berndt SI, The Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial, Bergmann S, Bennett AJ, Bingham SA, Bochud M, Brown M, Cauchi S, Connell JM, Cooper C, Davey Smith G, Day I, Dina C, De S, Dermitzakis ET, Doney ASD, Elliott KS, Elliott P, Evans DM, Farooqi IS, *et al.*: **Association studies involving over 90,000 people demonstrate that common variants near to *MC4R* influence fat mass, weight and risk of obesity.** *Nat Genet* 2008, **40:**768-775.
15. Bodmer W, Bonilla C: **Common and rare variants in multifactorial susceptibility to common diseases.** *Nat Genet* 2008, **40:**695-701.
16. Schork NJ, Murray SS, Frazer KA, Topol EJ: **Common vs. rare allele hypotheses for complex diseases.** *Curr Opin Genet Dev* 2009, **19:**212-219.
17. International Schizophrenia Consortium: **Rare chromosomal deletions and duplications increase risk of schizophrenia.** *Nature* 2008, **455:**237-241.
18. Stefansson H, Rujescu D, Cichon S, Pietiläinen OP, Ingason A, Steinberg S, Fossdal R, Sigurdsson E, Sigmundsson T, Buizer-Voskamp JE, Hansen T, Jakobsen KD, Muglia P, Francks C, Matthews PM, Gylfason A, Halldorsson BV, Gudbjartsson D, Thorgeirsson TE, Sigurdsson A, Jonasdottir A, Jonasdottir A, Bjornsson A, Mattiasdottir S, Blondal T, Haraldsson M, Magnusdottir BB, Giegling I, Möller HJ, Hartmann A, *et al.*: **Large recurrent microdeletions associated with schizophrenia.** *Nature* 2008, **455:**232-236.
19. Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, Erlich HA, Julier C, Morahan G, Nerup J, Nierras C, Plagnol V, Pociot F, Schuilenburg H, Smyth DJ, Stevens H, Todd JA, Walker NM, Rich SS; The Type 1 Diabetes Genetics Consortium: **Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes.** *Nat Genet* 2009, doi:10.1038/ng.381.
20. Russell RK, Nimmo ER, Satsangi J: **Molecular genetics of Crohn's disease.** *Curr Opin Genet Dev* 2004, **14:**264-270.
21. Ng MC, Park KS, Oh B, Tam CH, Cho YM, Shin HD, Lam VK, Ma RC, So WY, Cho YS, Kim HL, Lee HK, Chan JC, Cho NH: **Implication of genetic variants near *TCF7L2, SLC30A8, HHEX, CDKAL1, CDKN2A/B, IGF2BP2*, and *FTO* in type 2 diabetes and obesity in 6,719 Asians.** *Diabetes* 2008, **57:**2226-2233.
22. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA: **Rare variants of *IFIH1*, a gene implicated in antiviral responses, protect against type 1 diabetes.** *Science* 2009, **324:**387-389.
23. McCarthy MI, Hattersley AT: **Learning from molecular genetics: novel insights arising from the definition of genes for monogenic and type 2 diabetes.** *Diabetes* 2008, **57:**2889-2898.
24. Ghoussaini M, Song H, Koessler T, Al Olama AA, Kote-Jarai Z, Driver KE, Pooley KA, Ramus SJ, Kjaer SK, Hogdall E, DiCioccio RA, Whittemore AS, Gayther SA, Giles GG, Guy M, Edwards SM, Morrison J, Donovan JL, Hamdy FC, Dearnaley DP, Ardern-Jones AT, Hall AL, O'Brien LT, Gehr-Swain BN, Wilkinson RA, Brown PM, Hopper JL, Neal DE, Pharoah PD, Ponder BA, *et al.*: **Multiple loci with different cancer specificities within the 8q24 gene desert.** *J Natl Cancer Inst* 2008, **100:**962-966.

Published: 03 July 2009
doi:10.1186/gm66